MEDIAPART

SÉRIE EP. 1 L'INTELLIGENCE ARTIFICIELLE ET SON MONDE

L'intelligence artificielle, boîte noire au succès fulgurant et aux limites vertigineuses

Depuis le lancement de ChatGPT le 30 novembre 2022, l'IA générative s'impose à grande vitesse dans nos téléphones et nos ordinateurs. Ses applications potentielles sont légion, tout comme ses zones d'ombre. Cette technologie est déjà considérée comme une impasse par un de ses créateurs.

Dan Israel et Martine Orange - 9 février 2025 à 18h32

Le 30 novembre 2022 est une de ces dates qui ont tout chamboulé dans le rapport du grand public à la révolution numérique. Ce jour-là, l'entreprise OpenAI a rendu accessible à toutes et tous son robot conversationnel ChatGPT. En quelques mois, ce *chatbot* est devenu le symbole du succès de l'IA générative, qui s'impose toujours plus dans nos vies. Avec ses concurrents Llama, Gemini, Claude ou le petit nouveau chinois DeepSeek, ChatGPT donne le la d'une transformation explosive de nos rapports avec les machines.

Comment fonctionne cette technologie, dont les bases théoriques ont été jetées dès la fin des années 1950 ? Pourquoi produit-elle des erreurs de façon incompressible, poussant l'inventeur de sa forme moderne à la considérer d'ores et déjà comme une impasse ? Quels en sont les promesses ? et les dangers ? Tentative d'exploration des entrailles de la machine.

Des outils devenus omniprésents

En une poignée d'années, les programmes d'intelligence artificielle générative, capables de simuler le langage, le son et les images, se sont imposés dans notre vie numérique. Ils ont pris d'assaut les téléphones et les ordinateurs, pour un grand nombre d'usages : recherche et organisation d'informations, rédaction ou synthèse, aide aux devoirs, assistance administrative au sens large, mais aussi transcription écrite de documents sonores, traduction à la volée, génération automatique de sous-titres, création musicale de bonne facture, clonage de voix…

« *En cinq ou six ans, les progrès sont très impressionnants*, témoigne Morgan Blangeois, doctorant à l'université Clermont-Auvergne, qui prépare une thèse sur les bouleversements provoqués par l'IA générative dans le secteur des services numériques. *En matière de traduction, la machine sait désormais résoudre facilement le classique exemple du mot "bank", qui, en anglais, peut vouloir dire, selon le contexte, "banque" ou "rive" d'une rivière* » – tout comme elle ne se trompe plus sur les divers sens du mot « avocat ».

« *Un cas d'usage est bien établi : c'est l'assistance au codage, avec l'outil GitHub Copilot*, explique le jeune chercheur. *Il permet de compléter les actions du développeur, de débuguer, d'aller bien plus vite dans la production de codes. C'est déjà au point qu'un développeur m'a confié essayer de se "déperfuser" de cet outil.* »

« *Si on sait maîtriser les outils d'IA, dans certains domaines, ils sont déjà plus puissants que l'humain*, pointe Laurence Devillers, professeure à Sorbonne-Université/CNRS, chercheuse en intelligence artificielle et spécialiste des questions d'éthique dans ce domaine. *Si on soumet des millions de clichés de radiologie à une IA prédictive, la machine saura les interpréter, chercher des signaux faibles pour repérer des pathologies qui pourraient échapper à un radiologue.* »

Dans tous les domaines scientifiques, les applications possibles sont immenses. Ce n'est pas pour rien que le prix Nobel de Chimie 2024 vient d'être décerné à deux chercheurs de Google DeepMind, John Jumper et Demis Hassabis, pour le développement de leur outil AlphaFold, qui permet depuis 2020 de prédire la structure de protéines. Il y a quelques mois, un outil d'IA a aussi permis de déchiffrer une partie d'un papyrus complètement carbonisé lors de l'éruption du Vésuve en 79, à Herculanum, près de Pompéi.

Les déclinaisons à venir, dans tous les domaines, sont désormais innombrables. Et les inquiétudes des travailleurs et travailleuses qui craignent de se voir remplacé·es sont grandes. Le collectif de traducteurs est traductrices En chair et en os décrit par exemple un univers professionnel où l'arrivée en masse de ces nouveaux outils fait déjà de sérieux dégâts.

Des bases scientifiques posées depuis plus de soixante-cinq ans

Tenir dans sa poche un robot conversationnel parlant un français parfait (autant que l'anglais, l'allemand, le japonais…) est désormais une banalité, mais c'est le fruit d'une très longue marche. ChatGPT et ses concurrents de l'IA générative sont l'aboutissement du système des LLM (« *large language models* », « grands modèles de langage »), qui génèrent une phrase en se basant sur la plus grande probabilité statistique que tel mot aille derrière tel autre (« bleu » s'intègrera probablement après « le ciel est… »).

Nourris par des quantités colossales de données (récupérées sans trop s'embarrasser des droits de leurs autrices et auteurs originaux), les *chabtbots* obtiennent des résultats d'un très bon niveau pour imiter le langage courant et simuler des raisonnements ou des synthèses. Mais rien ne fonctionnerait sans les réseaux de neurones artificiels, à la base de ce modèle.

> « Tout cela était poussif : les ordinateurs peu puissants ne permettaient pas d'obtenir des résultats convaincants, et les recherches dans ce champ se sont arrêtées. »
>
> Jean-Gabriel Ganascia, professeur à Sorbonne-Université

Or, le premier réseau de neurones a été inventé en… 1958. « *Les concepts de base sont là très tôt*, sourit Jean-Gabriel Ganascia, professeur à la faculté des sciences de Sorbonne-Université, informaticien, spécialiste d'intelligence artificielle. *L'apprentissage par renforcement, base des méthode utilisées encore aujourd'hui, a été mis au point dès 1959, et l'apprentissage des réseaux de neurones date de 1986, il a même valu le prix Nobel 2024 à Geoffrey Hinton. Mais tout cela était poussif : les ordinateurs peu puissants ne permettaient pas d'obtenir des résultats convaincants, et les recherches dans ce champ se sont arrêtées.* »

La résurrection de ces méthodes, qui n'aurait pas eu lieu sans l'explosion de la puissance des ordinateurs et l'effondrement de leur coût, est due à un Français, le chercheur Yann Le Cun, aujourd'hui vice-président de Meta, qui pilote sa stratégie en matière d'IA et continue à faire de la recherche.

À partir de 2010, avec ses collègues Geoffrey Hinton et Yoshua Bengio, il remet au goût du jour « *l'apprentissage profond* », dont il est un des pionniers depuis les années 1970, en démontrant que cette méthode donne désormais d'excellents résultats. Toute la branche de l'IA le suit à partir de 2012-2013. En 2014, la génération d'images devient possible. En 2016, le logiciel Alphago (développé par la même société qu'AlphaFold) bat le Sud-Coréen Lee Sedol, un joueur de go légendaire.

Puis en 2017, Google invente le modèle « Transformer » (le « T » de ChatGPT), qui simplifie en profondeur les techniques de calcul nécessaires pour imiter le langage humain. L'année suivante, Google propose Bert, le premier LLM efficace, et Le Cun, Hinton et Bengio reçoivent le prix Turing, la plus haute récompense de l'informatique.

Une machine sans opinion ni conscience

En 2023, le prix Nobel Geoffrey Hinton a démissionné de son poste chez Google pour pouvoir relayer plus librement ses craintes au sujet de l'IA, dont il a contribué à créer la forme moderne. Au *New York Times*, il a dit son inquiétude devant l'avalanche d'infox et d'images truquées dont il anticipe le déferlement.

Dans une autre interview lors de sa remise de prix, il a dessiné un monde où la destruction d'emplois serait inévitable, face aux machines capables de traduire, de créer et de parler comme des humains. Surtout, il a exprimé son angoisse devant « *ces choses* » qui pourraient « *devenir incontrôlables et prendre le contrôle* » de la civilisation, nous enjoignant de « *trouver s'il existe une manière de faire face à cette menace* ».

Cette menace, régulièrement décrite par ceux-là mêmes qui contribuent au développement de l'IA, est aussi omniprésente dans la vulgate des grands patrons de la Silicon Valley, qui entretiennent en même temps le rêve transhumaniste d'un homme « augmenté » par la machine.

Tous guettent l'irruption d'une « IA forte », autrement dit l'avènement d'une machine dotée de conscience, prélude qui mènerait inéluctablement à « l'IA générale » et à la « singularité technologique » : le moment où l'intelligence des machines dépassera celle des humains, précipitant la chute de la civilisation humaine.

Tous les acteurs importants de la tech brandissent ce scénario ambivalent, entre ferveur millénariste et peur de l'apocalypse numérique. Et certains croient le voir advenir : en 2022, Blake Lemoine, un ingénieur de Google, est devenu célèbre pour avoir cru, en conversant avec un robot maison, discuter avec une entité intelligente, consciente et sensible.

> « Personne ne peut garantir que ce qui sort de la machine est factuel, non toxique, compréhensible. »
>
> Yann Le Cun, un des « pères » de l'IA générative

Pourtant, rien ne permet d'accréditer cette option aujourd'hui. « *Pour obtenir une machine consciente, il faut qu'elle puisse exprimer des émotions, des désirs. On en est loin. La technologie et les sciences actuelles ne le permettent absolument pas* », balaye Jean-Gabriel Ganascia, qui aime comparer les grands modèles de langage à « *des robots bavards* ». Des machines statistiques joliment déguisées.

« *Quant à la question de dépasser l'intelligence humains, cela demanderait que la machine accumule des facultés mentales, qu'on ne sait que simuler, et dont on n'a aucune idée de la manière dont elles pourraient s'accumuler. Cette idée est non fondée* », assure-t-il.

« *Nous projetons sur cette machine des capacités et des connaissances dont elle ne dispose pas*, estime elle aussi Laurence Devillers. *Elle ne fait qu'aligner des suites de mots en suivant notre prompt et nos intentions, sans intention et sans opinion propre.* »

Erreurs inévitables, biais dangereux

C'est le « père » de l'IA moderne qui en dit le plus de mal. À propos de l'intelligence artificielle générative, Yann Le Cun avait dès 2023 une formule expéditive et cruelle, mais éclairante : « *Personne ne peut garantir que ce qui sort de la machine est factuel, non toxique, compréhensible.* »

Le chercheur et vice-président de Meta estime que le modèle technologique des LLM est une impasse et enjoint à ses pairs de chercher d'autres voies. Lui-même travaille depuis deux ans sur un autre modèle, dont il entrevoit l'aboutissement dans seulement de très longues années.

Car tout le monde le sait, bien que presque personne ne le formule clairement : l'IA générative, qui aligne les mots statistiquement les plus probables, commet des erreurs, parfois beaucoup d'erreurs. Demander début février 2025 aux robots disponibles qui est le premier ministre français expose par exemple à se faire répondre qu'il s'agit d'Élisabeth Borne (ChatGPT et DeepSeek) ou Gabriel Attal (Claude). Ce sont les fameuses « hallucinations », un terme promu par les géants de la tech pour humaniser leurs créations et rendre acceptables leurs déraillements.

Selon l'entreprise de sécurité NewsGuard, qui soumet ces outils à des tests rigoureux, le taux d'échec moyen des dix principaux *chatbots* était de 62 % (d'erreurs ou de non-réponses) en décembre. Celui de DeepSeek un mois plus tard était de 83 %, et dans trois cas sur dix, le nouveau robot « *a relayé la position du gouvernement chinois sans qu'il lui ait été demandé quoi que ce soit concernant la Chine* ».

> La rapporteuse spéciale sur les formes contemporaines de racisme du Haut-Commissariat des droits de l'homme de l'ONU a rappelé l'été dernier que l'IA générative n'était ni neutre ni objective.

Les erreurs des machines dérivent bien souvent vers des biais racistes ou sexistes. Comme l'a résumé à Mediapart le journaliste technocritique Thibault Prévost, « *plus on s'éloigne de la médiane, plus le modèle va avoir du mal à prédire et à modéliser le monde* ». Le problème étant que « *la médiane politique du monde correspond à la bourgeoisie, à la blanchité, au genre masculin* ».

Dans une tribune parue dans *Le Monde*, un collectif d'ONG, dont Amnesty International et la Ligue des droits de l'homme, dénonce une IA qui « *perpétue les stéréotypes, renforce les inégalités sociales et limite l'accès aux ressources et opportunités* » des « *populations les plus*

vulnérables et les plus discriminées ». Les exemples sont déjà nombreux.

La rapporteuse spéciale sur les formes contemporaines de racisme du Haut-Commissariat des droits de l'homme de l'ONU, a rappelé l'été dernier que l'IA générative n'était ni neutre ni objective, donnant comme exemple les dangers de la « *police prédictive* », qui « *illustre bien la façon dont les préjugés raciaux sont reproduits par les avancées technologiques* ».

Une étude américaine a montré comment les modèles d'IA actuels évaluent négativement l'intelligence et l'employabilité des Noirs américains en se basant sur leur manière de parler. Une autre, menée à l'école de médecine de Stanford en Californie, a révélé que les préjugés racistes qu'ils véhiculent risquaient d'entraîner une prise en charge médicale dégradée pour les personnes victimes de ces stéréotypes. L'Unesco a aussi alerté sur les stéréotypes sexistes et homophobes de la plupart des LLM.

La non-transparence comme modèle

Le leader de l'utilisation grand public de l'IA générative, ChatGPT, ne correspond pas au nom de l'entreprise américaine qui l'a créé, OpenAI : il est tout sauf ouvert. « *Le système ChatGPT est entraîné sur des milliards de données qu'on ne connaît pas, et il est paramétré d'une manière que personne ne comprend avec certitude* », dénonce la chercheuse Laurence Devillers.

Celle-ci insiste sur la méconnaissance qu'a le grand public des concepts et des paramètres qui font fonctionner les *chatbots*. Qui a connaissance de la variable de « *température* », utilisée pour simuler la conversation humaine, en ajoutant une part de créativité et de hasard dans les réponses ? « *Si vous posez plusieurs fois la même question, vous n'obtiendrez pas la même réponse, et par conséquent, pas toujours la réponse la plus précise si la "température" est activée* », explique-t-elle.

> « OpenAI a une grande responsabilité pédagogique à l'égard de ses utilisateurs, dont il souhaite qu'ils soient un milliard cette année. »
>
> Morgan Blangeois, doctorant à l'université Clermont-Auvergne

Laurence Devillers souligne aussi la question des sources : « *Quand je crée un modèle à partir de milliards de données, je crée un puzzle, dont les éléments constitutifs ne peuvent plus être retrouvés, sauf si je l'interroge sur un sujet de niche : là, on peut retrouver un texte, et on peut parler de plagiat.* »

Au contraire, elle salue « *DeepSeek-R1, le système chinois qui a obtenu des performances supérieures à celles de ChatGPT 401* [le dernier modèle du robot américain – ndlr], *et qui montre que la voie la plus prometteuse est bien celle de l'*open source *et de la collaboration en recherche pour mieux maîtriser les systèmes* ». Les modèles de Meta sont également accessibles.

Pour le doctorant Morgan Blangeois, il est urgent d'ouvrir la boîte noire : « *OpenAI a une grande responsabilité pédagogique à l'égard de ses utilisateurs, dont il souhaite qu'ils soient un milliard cette année !* » Pour lui, « *il y a un très gros enjeu pédagogique à former et à informer la population, pour qu'elle sache utiliser l'outil, mais aussi le comprendre* ».

Une demande d'énergie insatiable

C'est la question qui taraude le secteur, mais qui reste soigneusement camouflée : va-t-il pouvoir trouver les moyens énergétiques suffisants pour répondre à ses besoins ? Selon le dernier rapport de l'Agence internationale de l'énergie, la consommation totale d'électricité liée au développement de l'intelligence artificielle, à commencer par l'énergie nécessaire aux *data centers*, devrait croître de 160 TWh en 2022 à 560 TWh en 2026 (un térawattheure équivaut à un milliard de kilowattheures). Rien qu'aux États-Unis, la consommation des *data centers* devrait doubler ces cinq prochaines années.

En quelques années, les entreprises du secteur ont investi dans l'intelligence artificielle des sommes dépassant les investissements dans le gaz et le pétrole, et promettent d'y consacrer encore plus d'argent. Selon le *Financial Times*, la somme se chiffre à 300 milliards de dollars rien que pour 2025.

Au Texas et en Californie, ce type d'installation totalise déjà plus de 10 % de la consommation totale. En Irlande, où les géants du numérique ont installé des hubs de *data centers* pour l'Europe, c'est 20 %.

Et plus la consommation augmente, plus les fragilités se révèlent dans le système électrique. Les moyens de production ne sont déjà plus toujours suffisants pour répondre à la demande, les populations installées aux alentours des centres de développement de l'IA subissant des parasitages et des chutes de tension.

Les géants du numérique sont « *dans une course pour la domination mondiale* », explique le président de Lancium, une société spécialisée dans l'installation de *data centers* au Texas. Le consultant Gartner estime que 40 % des data centers existants pourraient rencontrer des contraintes opérationnelles, en raison d'un accès limité à l'énergie.

Certains se sont donc rapprochés des producteurs d'électricité pour prendre des participations dans de nouvelles capacités de production. Google a ainsi récemment signé un accord de 20 milliards de dollars avec la société TPG Rise climate, pour développer des sites de production d'énergie près de ses centres. D'autres cherchent à se doter de centres de stockage d'énergie.

> Certains géants du numérique envisagent même de construire des miniréacteurs nucléaires – qui n'existent pas encore – à côté de leurs principaux centres de recherche et développement.

Dans la précipitation, certains fournisseurs ont décidé de rouvrir des centrales à charbon pour produire l'électricité voulue. Et en septembre, Microsoft a conclu un accord de vingt ans avec le groupe Constellation Energy pour relancer la centrale nucléaire de Three Mile Island. Alors qu'un des réacteurs avait été arrêté dès 1979 après le premier grave accident dans le nucléaire civil dans le monde, le second avait été stoppé en 2019.

Dans la foulée, Constellation Energy, persuadé que le développement de l'IA va permettre le relancement du nucléaire, a racheté son concurrent Calpine pour un peu plus de 26 milliards de dollars. Certains géants du numérique envisagent même de construire des miniréacteurs nucléaires – qui n'existent pas encore – à côté de leurs principaux centres de recherche et développement.

Alors qu'ils se présentaient encore récemment comme des champions de la lutte contre les dérèglements climatiques et de la transition écologique, tous ont enterré leurs bonnes résolutions : leurs installations d'énergies renouvelables – souvent des parcs solaires – ne vont pas suffire, de leur aveu même, pour répondre à leurs besoins. De même, ils ne mettent plus en avant les gains d'efficacité et de performance apportés par l'IA dans la gestion des réseaux électriques : les progrès réalisés risquent de ne pas égaler la hausse de la demande.

L'humain caché dans la machine

Invisibles et invisibilisé·es. Derrière les LLM et les succès de l'IA générative, on trouve des « petites mains » en masse, indispensables au triomphe des robots. C'est ce que martèle depuis près de dix ans Antonio Casilli, professeur de sociologie à Télécom Paris, auteur en 2019 de l'incontournable livre *En attendant les robots. Enquête sur le travail du clic* (Le Seuil). Il y établissait que qu'en matière d'IA, les failles de l'innovation sont nombreuses et doivent constamment être palliées par un recours intensif au travail humain, précarisé et délocalisé là où la main-d'œuvre est bon marché.

Il s'agit d'abord d'entraîner – le « P » de ChatGPT signifie « *pre-trained* » – et de fournir les machines en données exploitables. La matière première des robots est un « *mélange de stagiaires français et de précaires malgaches* », cinglait Antonio Casilli dans l'ouvrage, et il continue de le démontrer dans ses travaux.

Mediapart a en effet documenté comment de nombreuses entreprises francophones sous-traitent à Madagascar les tâches répétitives nécessaires à rendre « intelligents » leurs robots, pour améliorer la traduction, le sous-titrage automatique ou automatiser la perception d'une information par l'ordinateur.

On a aussi découvert qu'avant le déploiement en fanfare de ChatGPT, des Kenyans payés entre 1,3 et 2 dollars par jour ont été chargés, dans des conditions de travail apocalyptiques, de repérer les contenus « *toxiques* » sur le Web, obligés d'ingurgiter les pires textes et images du Net pour les épargner au client d'OpenAI.

Hommes et femmes doivent aussi assister, corriger, voire remplacer le robot. Quand Amazon a abandonné son

concept de magasin sans caisse, où les caméras étaient censées voir seules ce que le consommateur emportait avec lui, la supercherie a été dévoilée : pour moins de 200 magasins, il fallait un millier de travailleurs et de travailleuses indien·nes, chargé·es d’assurer la fiabilité du système.

Le *New York Times* a pour sa part bien documenté la manière dont les voitures autonomes nécessitent une supervision humaine à distance, évaluant celle-ci à 1,5 humain par véhicule autonome. Et cela devrait durer, tranche Jean-Gabriel Ganascia : « *Je ne pense pas qu’on pourra rendre une voiture réellement autonome en dehors d’une autoroute ou d’une ville américaine au plan très simple et où les piétons sont absents. Pour une ville européenne, remplie de piétons, je n’y crois pas. Les événements imprévus sont trop nombreux.* »

Dan Israel et Martine Orange

Le journal MEDIAPART est édité par la Société Editrice de Mediapart - 127 avenue Ledru-Rollin, 75011 Paris.
RCS Paris 500 631 932. Numéro de CPPAP : 1224Y90071 - Directeur de la publication : Carine Fouteau